MINING SPATIO-TEMPORAL INFORMATION SYSTEMS

Kluwer Academic Publishers

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
13-JUN-2002	Book Chapter (Referred)	

4. TITLE AND SUBTITLE
Spatio-Temporal Data Mining And Knowledge Discovery: Issues Overview

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0602435N

6. AUTHOR(S)
ROY VICTOR LADNER FREDERICK PETRY (Dr.)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
74-6731-02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Research Laboratory
Marine Geoscience Division
Stennis Space Center, MS 39529-5004

8. REPORTING ORGANIZATION REPORT NUMBER
NRL/BC/7440--02-1003

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release,distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Data mining or knowledge discovery refers to a variety of techniques having the intent of uncovering useful patterns and associations from large databases. The initial steps of data mining are concerned with preparation of data, including data cleaning intended to resolve errors and missing data and integration of data from multiple heterogeneous sources. Next are the steps needed to prepare for actual data mining including the selection of the specific data relevant to the task and the transformation of this data into a format required by the data mining approach. Finally specific data mining algorithms such as class description, association rules and classification clustering are applied. There are specific characteristics of spatial and temporal data, as found in GIS and multi-media data, that make knowledge discovery in this domain more complex than in mining ordinary data such as found in typical business sales applications. Here we provide a survey of work in spatio-temporal data mining emphasizing the special characteristics. An overview is given of different sources and types of geospatial, oceanographic and meteorological data and the associated issues inherent in their use in knowledge discovery.

15. SUBJECT TERMS
Data Mining, Spatio-temporal Data, Data Preparation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Roy Ladner
unclassified	unclassified	unclassified	Unlimited	19	19b. TELEPHONE NUMBER *(Include area code)* 228-688-4679

Chapter 1

Spatio-Temporal Data Mining and Knowledge Discovery: Issues Overview

Roy Ladner and Frederick Petry
Naval Research Laboratory, Stennis Space Center

Key words: Data Mining, Spatio-temporal Data, Data Preparation

Abstract: Data mining or knowledge discovery refers to a variety of techniques having the intent of uncovering useful patterns and associations from large databases. The initial steps of data mining are concerned with preparation of data, including data cleaning intended to resolve errors and missing data and integration of data from multiple heterogeneous sources. Next are the steps needed to prepare for actual data mining including the selection of the specific data relevant to the task and the transformation of this data into a format required by the data mining approach. Finally specific data mining algorithms such as class description, association rules and classification clustering are applied. There are specific characteristics of spatial and temporal data, as found in GIS and multi-media data, that make knowledge discovery in this domain more complex than in mining ordinary data such as found in typical business sales applications. Here we provide a survey of work in spatio-temporal data mining emphasizing the special characteristics. An overview is given of different sources and types of geospatial, oceanographic and meteorological data and the associated issues inherent in their use in knowledge discovery.

1. INTRODUCTION

Data mining or knowledge discovery generally refers to a variety of techniques that have developed in the fields of databases, machine learning and pattern recognition. The intent is to uncover useful patterns and associations from large databases. There are specific characteristics of spatial and temporal data, such as found in GIS and multi-media data, that make knowledge discovery in this domain more complex than in mining ordinary data such as found in typical business sales applications. In this chapter we

review briefly some background of data mining and specifically spatial data mining. Then we focus on some the issues that have arisen in our data mining research relative to spatial data characteristics that cause difficulties in data mining.

2. BACKGROUND

2.1 Data Mining

We shall first review the overall process of data mining. The initial steps of the process are concerned with preparation of data, including data cleaning intended to resolve errors and missing data and integration of data from multiple heterogeneous sources. Next are the steps needed to prepare for actual data mining. These include the selection of the specific data relevant to the task and the transformation of this data into a format required by the data mining approach. These steps are sometimes considered to be those in the development of a data warehouse, i.e., an organized format of data available for various data mining tools [Han and Kamber 2000]. There are a wide variety of specific knowledge discovery algorithms that have been developed [Hand et al. 2001]. These discover patterns that can then be evaluated based on some interestingness measure used to prune the huge number of available patterns. Finally as true for any decision aid system, an effective user interface with visualization / alternative representations must be developed for the presentation of the discovered knowledge.

Specific data mining algorithms can be considered as belonging to two categories - descriptive and predictive data mining. In the descriptive category are class description, association rules and classification. Class description can either provide a characterization or generalization of the data or comparisons between data classes to provide class discriminations. Association rules correspond to correlations among the data items and they are often expressed in rule form showing attribute-value conditions that commonly occur at the same time in some set of data. An association rule of the form X ➔ Y can be interpreted as meaning that the tuples in the database that satisfy the condition X also are "likely" to satisfy Y, so that the "likely" implies this is not a functional dependency in the formal database sense. Finally, a classification approach analyzes the training data (data whose class membership is known) and constructs a model for each class based on the features in the data. Commonly, the outputs generated are decision trees or sets of classification rules. These can be used both for the characterization of the classes of existing data and to allow the classification of data in the future, and so can also be considered predictive.

Predictive analysis is also a very developed area of data mining. One very common approach is clustering. Clustering analysis identifies the collections of data objects that are similar to each other. The similarity metric is often a distance function given by experts or appropriate users. A good clustering method produces high quality clusters to yield low inter-cluster similarity and high intra-cluster similarity. Prediction techniques are used to predict possible missing data values or distributions of values of some attributes in a set of objects. First, one must find the set of attributes relevant to the attribute of interest and then predict a distribution of values based on the set of data similar to the selected objects. There are a large variety of techniques used, including regression analysis, correlation analysis, genetic algorithms and neural networks to mention a few.

Finally, a particular case of predictive analysis is time-series analysis. This technique considers a large set of time-based data to discover regularities and interesting characteristics. One can search for similar sequences or subsequences, then mine sequential patterns, periodicities, trends and deviations.

2.2 Spatial Data Mining

There is now considerable interest in spatial data mining, but only recently have major research efforts been developed in this area. A major difference between data mining in ordinary relational databases and in spatial databases is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations), which are used by spatial data mining algorithms.

A very active and influential data mining research group is that led by Han in Vancouver and presently at Illinois. They have investigated several approaches to spatial data mining and have developed a system called GeoMiner [Han et al. 1997] based on these techniques. One approach developed a generalization-based knowledge discovery mechanism, which integrated attribute-oriented induction on non-spatial data and spatial merge and generalization on the spatial data [Lu et. al 1993]. The CLARANS clustering algorithm is a randomized search for an optimal cluster. Another spatial data mining approach was based on CLARANS and produced high-level non-spatial description of objects in every cluster using attribute-oriented induction [Ng and Han 1994].

One important topic for this area was the development of an approach for mining strong association rules in geographic information databases

[Koperski and Han 1995]. This approach uses an SQL-like spatial data mining query interface as developed for GeoMiner. This provides the subset of the spatial database over which the rule discovery is performed. From this subset the spatial predicates of interest such as intersect, adjacent, etc. are then explicitly materialized. The Apriori algorithm [Agrawal et al. 1993] is applied over this data to extract the association rules. If there is a concept hierarchy for the data and/or the spatial predicates, a multi-level approach to the Apriori algorithm allows rules to be extracted at any desired level.

A research group in Munich [Ester et al. 2000] has developed a set of database primitives for mining in spatial databases that are sufficient to express most of the algorithms for spatial data mining and that can be efficiently supported by a DBMS. They have found that the use of such database primitives enables the integration of spatial data mining with existing DBMSs and speeds-up the development of new spatial data mining algorithms. The database primitives are based on the concepts of neighborhood graphs and neighborhood paths. Effective filters allow restriction of the search to such neighborhood paths "leading away" from a starting object. Neighborhood indices materialize certain neighborhood graphs to support efficient processing of the database primitives by a DBMS. For spatial characterization it seems important that class membership of a database object is not only determined by its non-spatial attributes but also by the attributes of objects in its neighborhood. In spatial trend analysis, patterns of change of some non-spatial attributes in the neighborhood of a database object were determined. Spatial trends can be thought of as describing the regular change of non-spatial attributes when moving away from certain start objects for which both global and local trends can be distinguished.

Another approach has been taken by the use of spatial autocorrelation rather than materializing spatial predicates. The system is used to predict locations using map similarity [Chawla et al. 2000]. It has four components – map similarity measures, parametric functions for spatial models, a discretized parameter space and the search algorithm. The search explores the parameter space to discover the parameter value tuple maximizing the map similarity measure. These parameter values thus indicate the parametric function to use as the possible spatial model.

2.3 GIDB Data Mining

The setting in which we are developing approaches for spatial data mining is an environment whose objective is to develop ways of processing large amounts of spatio-temporal data especially of oceanographic and littoral regions and including meteorological information. Our goal is to integrate data mining techniques into the GIDB geospatial system described below.

The ultimate aim is to provide knowledge-enhanced information to decision tools that will be used by US Navy and Marine planners.

2.3.1 Data Mining Effort at NRL DMAP

We have applied several data mining techniques to spatial data of interest to the Navy. These include association rules, attribute generalization and predictive modelling. The predictive modelling is a support vector regression approach using a COTS system. We built a predictive model of wave heights and frequency using data from the twenty years of data observations of sea conditions at the Field Research Facility in Duck, North Carolina USA. The objective was to provide advisory information to tactical Naval planners for amphibious operations.

The attribute-oriented induction approach produces a generalized representation by either attribute removal or attribute generalization. We applied this technique to sea bottom data from 10 locations (such as areas in the Philippines, Mediterranean, Persian Gulf, etc.). Here the intended application was to characterize various sea bottom areas for the planning of a mine deployment / hunting mission. The spatial data was queried to formulate the files from which the attribute generalization was done. The basic query was on bottom sediment classification as this was the major characteristic of interest to experts.

Finally an extension of association rule discovery applied to fuzzy spatial data is being developed [Ladner et al. 2003]. Since the data we are interested in, as is typical of much spatial data [Burrough and Frank 1996], has uncertainty associated, we can model this using fuzzy sets [Cobb et al. 2000]. As an example consider using a spatial database to provide assistance in the logistical planning for a military operation. Then we might wish to uncover some of the important relationships of the data attributes in each area to provide guidance in the mission planning. An example of a possible rule that might be discovered is of the form:

If C is a small city and has good terrain nearby then there is a road nearby with 90% confidence.

Such a rule incorporates fuzzy information in the linguistic terms used such as "small" and "nearby."

2.3.2 Geospatial Information Database (GIDB™)

The Digital Mapping, Charting and Geodesy Analysis Program (DMAP) at the Naval Research Laboratory has been actively involved in the development of a digital geospatial mapping and analysis system since 1994. This work started with the Geospatial Information Database (GIDB™), an

object-oriented, CORBA-compliant spatial database capable of storing multiple data types from multiple sources. Data is accessible over the Internet via a Java Applet [Chung et al. 2001].

The GIDB includes an object-oriented data model, an object-oriented database management system (OODBMS) and various analysis tools. While the model provides the design of classes and hierarchies, the OODBMS provides an effective means of control and management of objects on disk such as locking, transaction control, etc. The OODBMS in use is Ozone, an open-source database management system. This has been beneficial in several aspects. Among these, access to the source code allows customization and there are no costly commercial database licensing fees on deployment. Spatial and temporal analysis tools include query interaction, multimedia support and map symbology support. Users can query the database by area-of-interest, time-of-interest, distance and attribute. For example, statistics and data plots can be generated to reflect wave height for a given span of time at an ocean sensor. Interfaces are implemented to afford compatibility with Arc/Info, Oracle 8i, Matlab, and others.

The object-oriented approach has been beneficial in dealing with complex spatial data, and it has also permitted integration of a variety of raster and vector data products in a common database. Some of the raster data include satellite and motion imagery, Compressed ARC Digitized Raster Graphics (CADRG), Controlled Image Base (CIB), jpeg and video. Vector data includes Vector Product Format (VPF) products from the National Imagery and Mapping Agency (NIMA), Shape, real-time and in-situ sensor data and Digital Terrain Elevation Data (DTED). The VPF data includes such NIMA products as Digital Nautical Chart (DNC), Vector Map (VMAP), Urban Vector Map (UVMAP), Digital Topographic Data Mission Specific Data Sets (DTOP MSDS), and Tactical Oceanographic Data (TOD).

Over the years, the system has been expanded to include a communications portal that enables users to obtain data from a variety of data providers distributed over the Internet in addition to the GIDB. These providers include Fleet Numerical Meteorology and Oceanography Center (FNMOC), USGS, Digital Earth/NASA, and the Geography Network/ESRI. A significant FNMOC product is the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) data. The atmospheric components of COAMPS are used operationally by the U.S. Navy for short-term numerical weather prediction for various regions around the world. Our communications gateway provides a convenient means for users to obtain COAMPS data and incorporate it with other vector and raster data in map form. The gateway establishes a well-defined interface that brings together such heterogeneous data for a common geo-referenced presentation to the user. An illustration of the interface for a typical data request is shown in Figure 1.

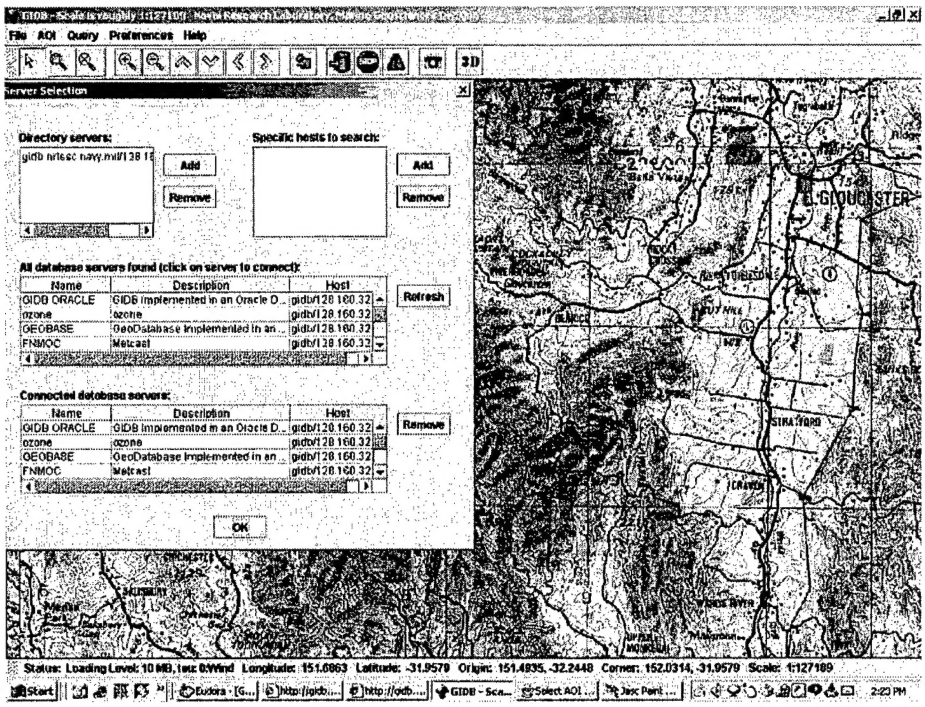

Figure 1. GIDB Interface

3. DATA

This section surveys different sources and types of geospatial, oceanographic and meteorological data and associated issues inherent in their use in knowledge discovery. Some data sources include the National Imagery and Mapping Agency (NIMA), Naval Oceanographic Office (NAVO) and FRF. Our overview of these spatio-temporal data sources may be of use to research in data mining as some of these are sources not commonly known outside the DOD community. Such a diverse mix of data will be seen to present a number of troublesome issues for knowledge discovery.

3.1 Vector Data - NIMA

The National Imagery and Mapping Agency (NIMA) is a major source of environmental data for the Department of Defense and the private sector. In the 1980's NIMA began the process of transforming their paper mapping data to digital format with a new database specification, Vector Product Format (VPF). Generally, VPF separates data into thematic coverages, with each of

these coverages containing thematically consistent data [VPF 96]. More details on VPF are given below.

A detailed listing of NIMA's digital data is available in [NIMA]. Table 1 lists NIMA's VPF products. Each product is designed to fill different needs. Digital Nautical Chart (DNC) for example, is directed at marine navigation and GIS applications, and it contains significant features collected from harbor, approach, coastal and general charts. Digital Topographic Data (DTOP) is produced for specific geographic areas. DTOP consists of thematic layers from terrain analysis and topographic line maps. Themes include vegetation, transportation, surface materials, surface drainage, obstacles, surface configuration or slope, hydrography, boundaries, population, industry, physiography, utilities and data quality. Mission Specific Data Sets (MSDS) are produced by NIMA in preparation for specific military missions.

Table 1. Partial Listing of NIMA's VPF Products

Product Name	Abbreviation
Digital Nautical Chart	DNC
Digital Topographic Data	DTOP
Vector Map	VMAP
Urban Vector Map	UVMAP
World Vector Shoreline	WVS
Tactical Terrain Data	TTD
Foundation Feature Data	FFD
Mission Specific Data Sets	MSDS
Interim Terrain Data	VITD

Tactical Terrain Data (TTD), consisting of DNC, DTOP and Digital Terrain Elevation Data (described below), is intended to provide data critical to planning and executing joint operations such as close air support missions, amphibious operations and land combat operations. TTD is supportive of applications that are to be used for terrain visualization, mobility planning, site and route selection, reconnaissance and communications planning, navigation and munitions guidance. TTD data density is generally consistent with similar portrayals on topographic line maps, terrain analysis products and hydrographic charts.

Interim Terrain Data (ITD) was designed to provide digital terrain analysis data for systems fielded before the production of Tactical Terrain Data. It consists of six thematic coverages or layers: vegetation, surface material, surface slope, surface drainage, obstacles and transportation. Features correspond to a 1:50,000 scale map [NIMA, USAS 98].

Vector Map (VMAP) is provided in Levels 0, 1 and 2, each increasing from small to large scale. Data coverages include boundaries, elevation, hydrography, industry, physiography, population, transportation, utilities, and vegetation. Urban Vector Map (UVMAP) provides specific vector-based

geospatial data with city graphic content. The same coverages are provided as for VMAP. Detail is similar to NIMA city graphic and military city map products.

World Vector Shoreline (WVS) content includes shoreline, international boundaries, maritime boundaries and country labels. Five libraries provide data derive from 1:250,000 to 1:12,000,000 scale source.

Digital Feature Analysis Data (DFAD) is a source of digital feature data. DFAD feature data is assigned an identification code and is described in terms of height, composition, length and orientation. Data is stored in vector format with one record for each feature. Each record contains coded attributes and a coordinate string. DFAD is collected from photogrammetric as well as cartographic source material. DFAD Level 1 offers medium scale detail (1:250,000) and Level 2 offers higher scale (1:50,000). The types of features included in DFAD include roads, railways, drainage, prominent buildings in urban areas, and prominent towers and power lines.

3.2 Miscellaneous Data Repositories

Data repositories and clearing houses hold spatial environmental data in a centralized location or make the availability of data at different locations known to prospective users. One such repository is the Terrain Resource Repository (TRR). NIMA's Terrain Modeling Project Office (TMPO) maintains the TRR. The TRR provides Internet access to various terrain data products available from the Department of Defense [TMPO 2000]. Users can access samples of many of NIMA's standard data products, along with software for viewing. The TRR also provides links to numerous web sites that are sources for environmental data. Among these are data sources maintained by state agencies, the U.S. Geologic Survey, the National Oceanographic and Atmospheric Administration, the U.S. Census Bureau, the U.S. Department of Transportation, the U.S. Department of Agriculture, the Bureau of Land Management, the Canadian Government, and the United Nations.

The Master Environmental Library (MEL) is another repository. The Defense Modelling Simulation Office (DMSO) maintains MEL. MEL indexes environmental data source location. Through MEL, users can locate and order environmental data online [MEL 2000].

The Tactical Oceanography Wide Area Network (TOWAN) is also a potential data repository. The Naval Research Laboratory provides TOWAN as an online environmental data repository and server that allows Department of Defense personnel and their contractors to search for and retrieve environmental information. TOWAN makes oceanographic databases

available, including bathymetric, geoacoustics, ice and magnetics. TOWAN is one of the nodes in MEL [TOWAN 2000].

The National Geospatial Data Clearinghouse (NGDC) aggregates numerous spatial data servers and provides a search interface. Search options include location, time period of content, full text and fielded search using country names or U.S. placenames. A custom search allows users to define parameters including map, temporal and server [NGDC 2000].

The National Oceanographic and Atmospheric Administration Server (NOAA) provides an on-line search by area-of-interest access to several databases. These databases include the NOAA Central Library, the Japan Science and Technology Corporation, the Foreign Data Library, the Office of Oceanographic and Atmospheric Research, the National Weather Service, and the National Snow and Ice Data Center [NOAA 2000].

While centralized data repositories and clearinghouses make the existence of data known to prospective users, they do not necessarily provide a means to actually use the data once it is obtained. There are numerous data models in which spatial data are organized and many proprietary database formats. Whether the data is acquired from a resource provider such as NIMA, a repository or another producer, it entails learning another data model and data format and also means writing code to import into the user's native format.

3.3 Oceanographic Data

The Naval Oceanographic Office (NAVO) produces and processes oceanographic data for the US Navy. Among the data produced by NAVO relevant to our study is bottom sediment data produced for mine warfare operations. This covers various locations worldwide and consists of polygons attributed with bottom sediment type (i.e., sand, coral, clay, gravel, silty sand, etc.).

3.4 Model Output

Fleet Numerical Meteorology and Oceanography Center (FNMOC) runs atmospheric predictive models [FNMOC 2002]. Representative of the nature of the output is the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) data. The atmospheric components of COAMPS are used operationally by the U.S. Navy for short-term numerical weather prediction for various regions around the world. COAMPS output includes data about expected precipitation, evaporation, winds, humidity, temperature, dew point depression, etc. Each parameter is given for various atmospheric levels, usually at 3-hour increments for a 72-hour period. This is gridded data over specific areas of the earth, with some products providing worldwide coverage.

The Air Force Weather Agency and the US Weather Service produce similar model output.

3.5 Observational Data – Argus Sites

There are twelve Argus Study Sites located at various locations around the world that record and archive observations of sea and atmospheric conditions [Argus 2002]. Sensors record changing waves, winds, tides and currents on approximately an hourly basis. Imagery makes possible the ascertainment of sand bar locations. The sensor data together with the imagery opens the door to analysis of factors influencing beach morphology. One such site is maintained at the army Corps of Engineers Field Research Facility (FRF) at Duck, NC. This is detailed data that can be a valued source for knowledge discovery.

3.6 2.5D and 3D Data

Bathymetric data is found in the Digital Bathymetric Data Base (DBDB). DBDB is gridded data giving ocean depths in meters for each 5 minutes of latitude and longitude worldwide. A primary source of terrain elevation data is NIMA's Digital Terrain Elevation Data (DTED). DTED comes in several resolutions ranging from 100 meter (Level 1) to 30 meter (Level 2) and down to 1 meter (Level 5). DTED is formatted in a uniform matrix of terrain elevation values, in 1° by 1° cells identified by southwest corner coordinates. DTED can be used to determine landform, slope, elevation or gross terrain roughness.

Acquisition of two-dimensional data is well developed in terms of NIMA's conversion of its paper map products to digital format and in the areas of extraction of feature data and three-dimensional elevation data from imagery. Satellite imagery is publicly available at one-meter resolution. DTED is available at resolutions of at least one meter from stereo pairs. The acquisition of three-dimensional geometric data related to man-made features is much more difficult. More geometry is needed to reconstruct features as three-dimensional objects rather than merely represent them symbolically as flat polygonal features or line features. Yet other objects may occlude much of the geometry that needs to be extracted from aerial and satellite imagery. Methods for acquiring such data include the automated extraction from imagery using photogrammetric techniques with high-resolution imagery, panoramic urban photographs and digital video imagery. [Irvin 89, Roux 94, Ernst 2000, Geometrix 2000].

4. DATA ISSUES

Data preparation is a crucial step to effective data mining. Data preparation in the spatial-temporal data context involves not only classic data 'cleaning' but also issues peculiar to the nature of spatial and temporal data. The latter include resolution of coordinate systems, datums and scale for spatial data, and resolution of temporal synchronicity and granularity for temporal data. This section presents an overview of these matters followed by issues encountered in data preparation in relation to specific data.

4.1 Spatio-Temporal Data Issues

By its very nature, spatial data exists in a coordinate reference system that locates the data somewhere on the earth. Generally, the mathematical model of the earth used to calculate coordinates is called a datum. Spatial data preparation includes ensuring that all data are on the same datum. This is important because coordinates for a point on the earth's surface calculated on one datum will not match the coordinates for the same point calculated on another datum. Many local datums have been developed over the years to satisfy mapping requirements for specific regions. Examples of these are the North American Datum of 1927 (NAD27), the European Datum and the more obscure Afgooye Datum applicable to Somalia. The World Geodetic System 1984 (WGS84) has been developed to provide a unified world system for expressing geodetic data. WGS84 is the datum currently in use by the US Department of Defense, and NIMA products are based on the WGS84. Public transformation parameters exist for conversion of coordinates from localized datums to WGS84 [NIMA 2000].

Just as it is important to ensure that all data is on the same horizontal datum, it is equally important to verify that all vertical measurements are on the same vertical datum, when vertical measurements are important to the knowledge discovery task. Use of mean sea level (MSL) is common, but many others can also be found such as low water, mean lower low water, neap tide, or even as height above the ellipsoid, to name just a few. Most NIMA products express elevation in terms of MSL. Global positioning systems give height above the ellipsoid.

In addition to datum issues, working with spatial data in knowledge discovery requires an understanding of the data's coordinate reference system. Rules for measuring distance, for example, in one coordinate system may not hold true for another. Even more basic is the consideration that all data should be on the same coordinate system in order to simplify spatial analysis. The location of spatial data is often expressed with geodetic coordinates of latitude and longitude. These are based on an ellipsoid model of the earth. Latitude is the angle between the plane of the equator and a line perpendicular

to the ellipsoid at a given point. Longitude is measured with respect to a prime meridian, positive to the east and negative to the west. Knowledge discovery in spatial data often involves spatial nearness or distance relationships. Geodetic coordinates, however, are non-Cartesian and require a more complicated mathematical calculation, great circle path, to measure distance. [Snyder 1987]

Figure 2 shows a screenshot from the GIDB application. The map shows NIMA DNC with NAVO bottom sediment data for the Onslow Bay area near the coast of North Carolina. The data is shown overlaid on a CADRG, a raster map produced by NIMA. Any datum and coordinate system incompatibilities are resolved at the time the data are imported into the GIDB database.

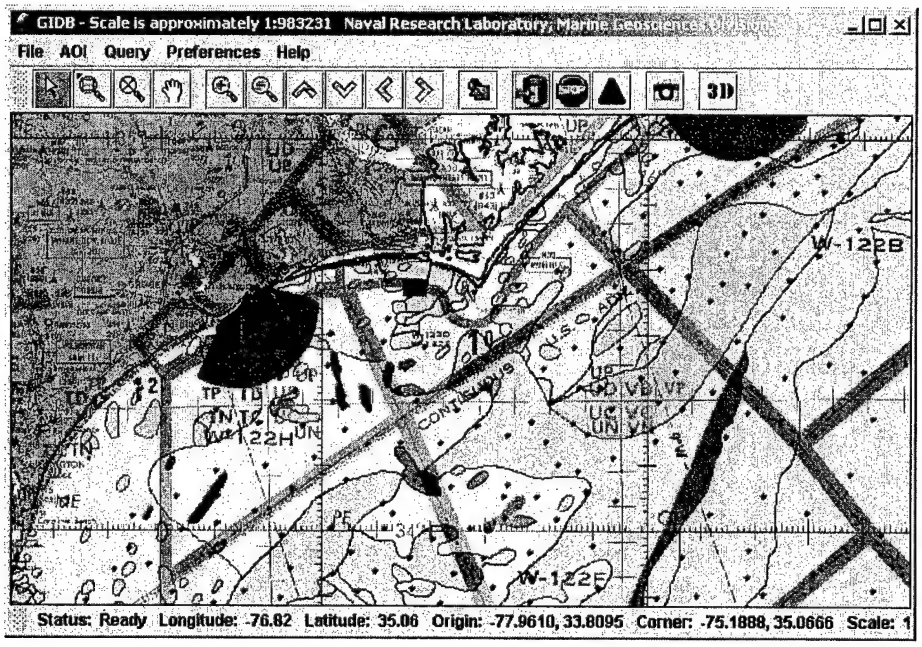

Figure 2. GIDB Screenshot Showing NIMA DNC with NAVO Bottom Sediments

In addition to geodetic coordinates, the location of spatial data may also be expressed in projected coordinates. Map projections are common in map making. They minimize some distortion while introducing others. Some common projections include the Mercator, the Transverse Mercator (TM), Universal Transverse Mercator (UTM) and the Lambert Conformal Conic (LCC). The Mercator is often used in navigation, the TM in quadrangle maps, the UTM in large-scale military maps, and the LCC in maps with predominantly east-west expanse [Snyder 1987]. Some projections may make

analysis of spatial relationships easier while other may introduce spatial distortions. Those that preserve area may help with analysis of spatial extent, while those that preserve scale and direction may help with analysis of nearness. Data preparation may involve conversion of all coordinates to a single projection or conversion to geodetic, according to the needs of the user.

Figure 3 shows the same area as figure 2 with the additional overlay of gridded and contoured weather data from FNMOC. The wind barbs show model output wind data and the contour lines illustrate predicted temperatures. The GIDB application allows the user to query these vector features in order to determine wind and temperature attribute values. The same holds true for any vector surface features such as bottom sediment type or sounding depth.

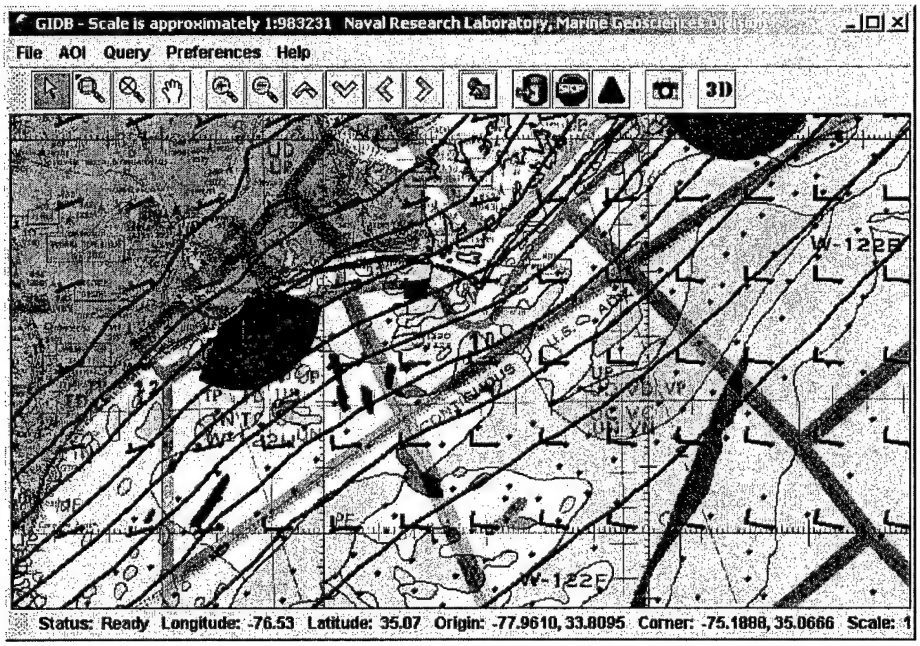

Figure 3. GIDB Screenshot Showing the Additional Overlay of Weather Data from FNMOC

Much digital spatial data, especially that from NIMA, is derived from digitized paper maps. Addressing map-scale and especially map-scale differences among such data is also an important step of the data preparation phase of mining spatial data. Large-scale (such as 1:20,000) spatial data will likely yield more feature-rich detail. Small-scale (such as 1:2,000,000) data may not. Among the questions that need to be answered are whether the data has sufficient detail to yield results. Small-scale data may also contain abstract representations of significant features or omit them entirely [USGS 2002]. A map of a winding river at one scale may be represented as a nearly

straight line at another and completely omitted at yet another. Whether the corresponding digital spatial data is an accurate representation of reality or an abstraction may strongly affect the results of the knowledge discovery process. Combining data from multiple scales can also impact spatial reasoning operations.

The potential effect of scale differences is shown in Figure 4. Two maps are presented over the Mount Rainier area but at two very different scales. Detailed contour line data at the mountain crest shown in the map on the left at 1:24,000 scale, is clearly missing from the 1:100,000 scale-map on the right. Corresponding digital data, if captured by digitization of the 1:100,000-scale map, would also omit the detailed contour data shown in the 1:24,000-scale map.

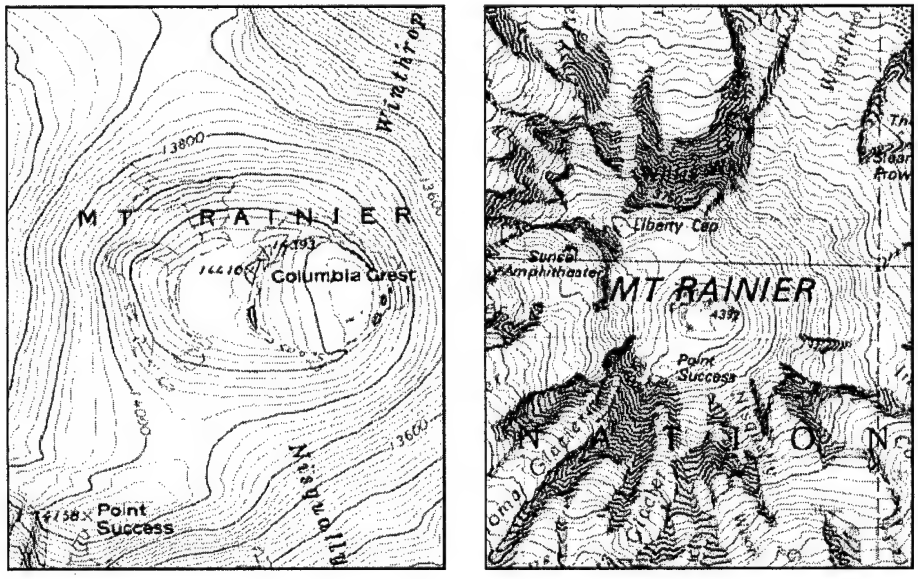

Figure 4. 1:24,000 Scale-Map on Left Compared with 1:100,000 Scale-Map on Right for Same Area [USGS 2002]

Temporal data present similar issues involving temporal synchronicity and granularity. Consider, for example, temporal atmospheric data consisting of measurements of winds, atmospheric pressure, rainfall, etc., each measured by different sensors that record events at various time intervals. Some measure each hour, some every six minutes, some once per day, some once every month, etc. Depending on the needs of the user, data preparation may require some combination of interpolation and binning in order to relate all atmospheric events to precise time intervals.

4.2 Data Source Issues

4.2.1 NIMA data

Most vector data distributed by NIMA is in Vector Product Format (VPF). A VPF database is made up of libraries that aggregate data by scale. Libraries are organized into coverages of thematically consistent data that share a single coordinate system and scale and that are contained within a specified spatial extent. Coverages can be tiled or geographically subdivided for the purpose of improving data management. A tiling scheme defines tile boundaries, the size of tiles and the handling of the features that lie on tile boundaries and text primitives that cross boundaries. Tiling schemes are defined by product specifications rather than by VPF. Coverages are composed of features. Five categories of cartographic features are defined in VPF: Point, Line, Area, Complex and Text. The coordinate system is defined at the library level. [VPF 1996]

This data organization generally follows historic mapping techniques of organizing data in disjoint thematic layers. While this method is favored for the production of maps in which layers are overlaid to produce the desired map view of the world, its usefulness for other purposes can be problematic. In particular, one study [Trott 1996] found that this practice resulted in considerable preprocessing and data integration to resolve inconsistencies between layers (such as positional errors) and to resolve missing attribute and geometric data for purposes of 3D synthetic environment construction.

For example, in some cases in the Digital Nautical Chart, feature attributes may be limited to the Digital Geographic Information Exchange Standard's (DIGEST) general feature and attribute coding catalogue (FACC) codes. While this provides an adequate description of what a feature is on a map, it fails to give detailed values with respect to this particular feature. In some cases, attribute values are designated as 'unknown.' NIMA's Mission Specific Data Sets (MSDS) on the other hand can be a good source of detailed data for a given area, with populated attributes. Since these are prepared in relation to specific missions, they may not exist for a geographic area of interest. Of note is the practice of some VPF products to aggregate varying scales within a database's library resulting in a scale-range rather than uniform scale within the library. For knowledge discovery this means adapting the appropriate knowledge discovery technique to the data.

4.2.2 Model output

This section covers issues relevant to use of gridded model output of the nature of that produced by FNMOC as discussed above. This data covers a spatial extent but is highly dependent on the temporal element in that it

consists of atmospheric forecasts at short-term periodic intervals originating from the time of the model run. Format is GRIB, a WMO (World Meteorological Organization) standard format for archiving and exchanging gridded data. It is a binary format that offers some data compression. Suitable software is necessary to read the file. Because of the data volume of model output, little historic output is warehoused, necessitating a user-created data store for examining long-term trends or for examining predicted conditions at some time in the past. Further, coordinates for the gridded data values are projected in the Mercator projection. Care must be taken when merging this data with other spatial data to assure a common coordinate system.

4.2.3 Observation data

This section discusses issues concerning observation data in the context of the data collected at the Argus Site at the FRF mentioned above. The discussion should be pertinent to a wide range of sensor-based observation data and serves to highlight areas in which special sensitivities are required in the data prep stage of knowledge discovery.

As discussed above, this data consists of detailed sensor readings over long periods of time. Its format is plain text and data for the Argus Site at the FRF is available on the Internet. Sensors are periodically replaced and others temporarily malfunction or are turned off for maintenance. This effectively produces "holes" in which observations are missing in recorded series. In addition, sensors record at widely differing intervals ranging from minutes to days or more. Of particular note is that the data provided to the public does not always differentiate between actual readings and a time-series summary. Hourly reports of wave heights, for example, may be a summary of the existing conditions over a one-hour time period. Further, the reported wave height may not have been "observed" but derived from pressure measurements.

Spatial considerations of datum and projection are relevant here. The spatial locations of sensors and physical measurements of sand bar locations, etc. at the FRF site, for example, are reported in a local coordinate system. That local system uses cross-shore and along-shore axes to designate x,y coordinates. Conversion to latitude and longitude or some other common projection is necessary in order to relate this data to other spatial data.

5. CONCLUSION

In this chapter we have provided an overview of the issues relevant to spatio-temporal data mining and knowledge discovery. We reviewed some background of data mining and specifically spatial data mining. Then we focused on some of the issues that have arisen in our data mining research relative to spatial data characteristics that cause difficulties in data mining. We described several of the sources of geospatial, oceanographic and meteorological data and outlined some of the specific characteristics of this type of data that make knowledge discovery in this domain more complex than in mining data such as found in typical business sales applications.

6. ACKNOWLDEGMENTS

We would like to thank the Naval Research Laboratory's Base Program, Program Element No. 0602435N for sponsoring this research.

Agrawal R Imielinski T and Swami A Mining Association Rules between sets of items in large databases. In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data.* New York, NY, ACM Press:207-216, 1993.

ARGUS, http://cil-www.oce.orst.edu:8080/.

Burrough P and Frank A (eds.) *Geographic Objects with Indeterminate Boundaries,* GISDATA Series Vol. 2, London, UK , Taylor and Francis, 1996.

Chawla S Shekar S Wu W and Ozesmi U Predicting Locations using Map Similarity (Plums): A Framework for Spatial Data Mining. In *Proceedings 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, ACM Press : 243-251, 2000.

Chung M Wilson R Cobb M Petry F and Shaw K Querying Multiple Data Sources via an Object-oriented Spatial Query Interface and Framework. *Journal of Visual Languages and Computing* **12**: 37-60, 2001.

Chung M Wilson R Ladner R Lovitt T Cobb M Abdelguerfi M and Shaw K The Geospatial Information Distribution System (GIDS). In Chaudhri A and Zicari R (eds) *Succeeding with Object Databases.* New York, NY, Wiley and Sons: 357-378, 2001.

Cobb M Petry F and. Robinson V Special Issue: Uncertainty in Geographic Information Systems and Spatial Data, *Fuzzy Sets and Systems*, **113**, #1, 2000.

Ernst I *3D City -Adaptive Capture and Visualization of Cityscapes*, GMD First, German National Research Center for Information Technology Institute for Computer Architecture and Software Technology, http://www.first.gmd.de/vista/3dcity/, January 2000.

Ester M Fromelt A Kriegel H and Sander J Spatial Data Mining: Database Primitives, Algorithms and Efficient and DBMS Support. *Data Mining and Knowledge Discovery* **4**: 89-125, 2000.

FNMOC, http://www.fnmoc.navy.mil/, April 2002.

Geometrix, Inc., http://www.geometrixinc.com/, January 2000.

Han J and Kamber M *Data Mining: Concepts and Techniques.* San Diego, CA Academic Press, 2000.

Han J Koperski K and Stefanovic N GeoMiner: A system prototype for spatial data mining. In *Proceedings of the 1997 ACM-SIGMOD International Conference on Management of Data*. New York, NY, ACM Press : 553-556, 1997.

Hand D Mannila H and Smyth P *Principles of Data Mining*. Cambridge, MA MIT Press, 2001.

Irvin, R. Bruce, and David M. McKeown, Jr., *Methods for Exploiting the Relationship Between Buildings and Their Shadows in Aerial Imagery*, IEEE Transactions on Systems, Man, and Cybernetics, **19**, No. 6, 1989.

Koperski K and Han J Discovery of Spatial Association Rules in Geographic Information Databases. In *Proceedings of 4th International Symposium on Large Spatial Databases.* Berlin, GD, Springer-Verlag: 47-66, 1995.

Ladner R Petry F Cobb M Fuzzy Set Approaches to Spatial Data Mining of Association Rules.To appear in *Transactions in Geographic Information Systems*, 2003.

Lu W Han J and Ooi B Discovery of general knowledge in large spatial databases. In *Proceedings of Far East Workshop Geographic Information Systems*. Singapore, World Scientific Press: 275-289, 1993.

MEL, Master Environment Library, Defense Modeling & Simulation Office, http://mel.dmso.mil, January 2000.

Ng R. and Han J Efficient and effective clustering method for spatial data mining.In *Proceedings of 1994 International Conference on Very Large Database*. San Francisco, CA, Morgan Kaufmann: 144-155, 1994.

NGDC, Federal Geographic Data Committee (FGDC) National Geospatial Data Clearinghouse (NGDC), http://www.fgdc.gov/ http://130.11.52.178/gateways.html], January 2000.

NIMA, *Digitizing the Future*, National Imagery and Mapping Agency.

NIMA 2000, National Imagery and Mapping Agency, Technical Report 8350.2, January 3, 2000.

NOAA, National Oceanographic and Atmospheric Administration, http://www.esdim.noaa.gov/noaaserver-bin/NOAAServer, January 2000.

Roux, Michel, and David M. McKeown, Feature Matching for Building Extraction from Multiple Views, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 46-53, 1994.

Snyder J *Map Projections – A Working Manual*, U.S. Geological Survey Professional Paper 1395, U.S. Government Printing Office, Washington, D.C., 1987.

TMPO, Terrain Resource Repository, Terrain Modeling Project Office, http://www.tmpo.nima.mil/mel, January 2000.

TOWAN, Tactical Oceanography Wide Area Network, Naval Research Laboratory, Stennis Space Center, http://www7180.nrlssc.navy.mil/homepages/TOWAN/TOWAN.htm, January 2000.

Trott K *Analysis of Digital Topographic Data Issues in Support of Synthetic Environment Terrain Data Base Generation*, TEC-0091, U.S. Army Corps of Engineers, Topographic Engineering Center, November 1996.

USAS, U.S. Army Simulation, Training, and Instrumentation Command, Orlando, Florida, *SEDRIS and The Synthetic Environment Domain, Volume 1 of the SEDRIS Document SET*, 12350 Research Parkway, Orlando, FL, March 28, 1998.

USGS, Map Scales, Fact Sheet, 015-02, U.S. Geological Survey, February 2002.

VPF, Departmentment of Defense, *Interface Standard for Vector Product Format*, MIL-STD 2407, 28 June 1996.